**Problem Statement:**

Below is the excel dataset, which includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.



We have to build a model that can predict whether the loan of the applicant will be approved or not based on the details provided in above excel dataset.

**Data Analysis:**

I have analysed the length of the Data is 614 \* 13. I found that, there are data with object type, so I must convert object type data into numerical values. I also found that there are some null values, so updated those null values by using “Mean” and “Mode” method.

**EDA:**

To Convert the Object data into numeric used below commands.

from sklearn.preprocessing import LabelEncoder

lab\_enc = LabelEncoder()

data1 = lab\_enc.fit\_transform(data['Loan\_ID'])

data2 = lab\_enc.fit\_transform(data['Gender'])

data3 = lab\_enc.fit\_transform(data['Married'])

data4 = lab\_enc.fit\_transform(data['Dependents'])

data5 = lab\_enc.fit\_transform(data['Education'])

data6 = lab\_enc.fit\_transform(data['Self\_Employed'])

data7 = lab\_enc.fit\_transform(data['Property\_Area'])

data8 = lab\_enc.fit\_transform(data['Loan\_Status'])

data['Loan\_ID'] = data1

data['Gender'] = data2

data['Married'] = data3

data['Dependents'] = data4

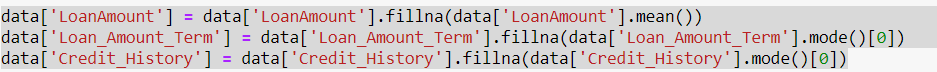
data['Education'] = data5

data['Self\_Employed'] = data6

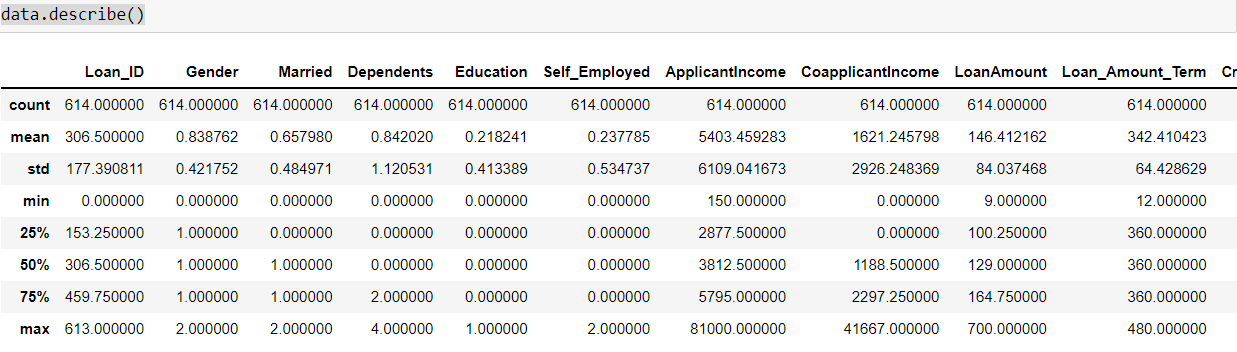
data['Property\_Area'] = data7

data['Loan\_Status'] = data8

Now used “Mean” and “Mode” method to replace null value with some numeric value.



Now I need to check the total health of my Dataset by using “.describe” method.



Now 1st of all I have to check the count and it has to be equal for all the attributes. I found that the count of all the attributes in above table is same.

Now next step is to check “mean” and “std” of continuous features (Ex:- ApplicantIncome, LoanAmount) and “std” should be lower than “mean”. If “std” is more, it means that data is deviating.

Next, I have to check “Quantile Range”, Which has to be equal, but I found in some feature it is high. So, it means there are some “Outliers” available in data set which I have to remove.

Next, I have to check data distribution of our dataset with the help of plotting. I need to see our continuous features of dataset whether it is equally distributed or not.

Text

Description automatically generated

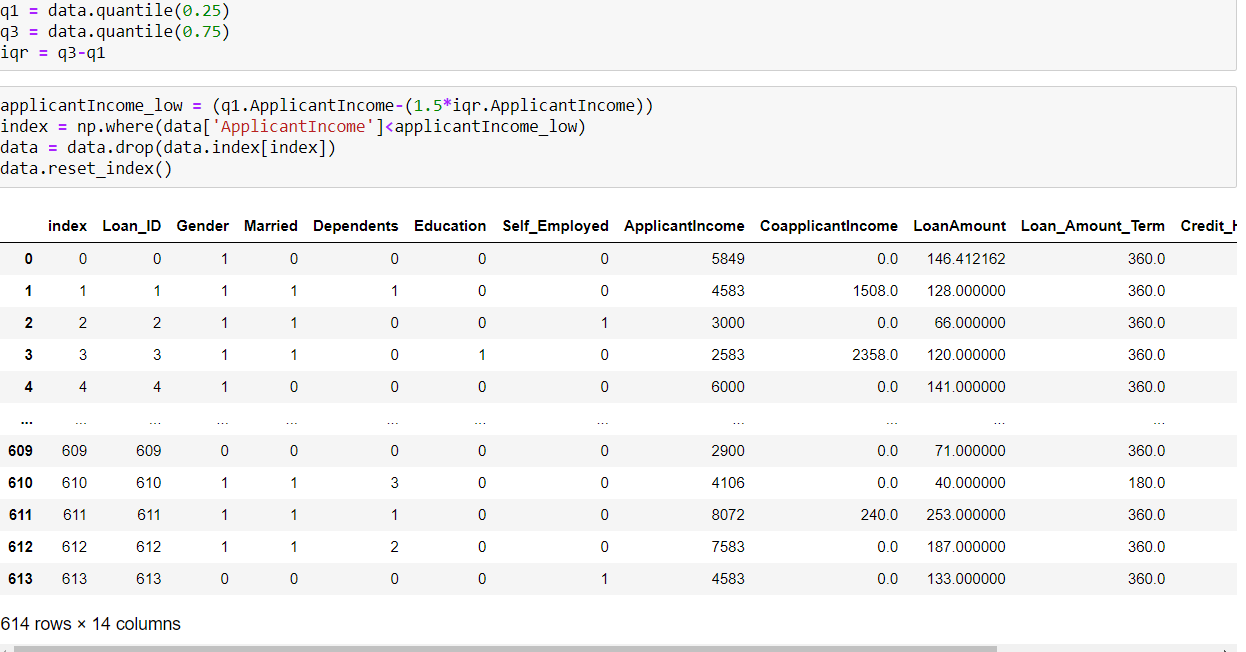
Graphical user interface, chart, histogram

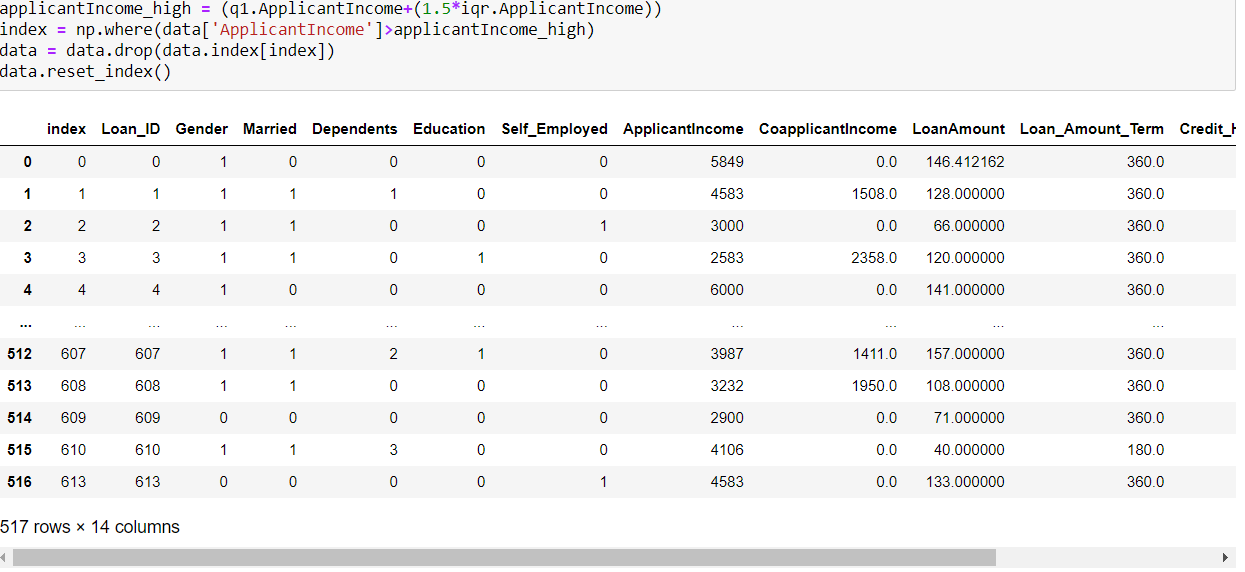
Description automatically generated

Chart, histogram

Description automatically generated

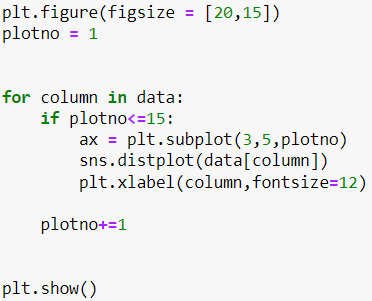
If it is a bell-shaped curve, it means there is no “Outliers” available but, in above plotting, I can see some of the continuous features is skewed, it means, it has some outliers which I have to remove with help of “Quantile method”.

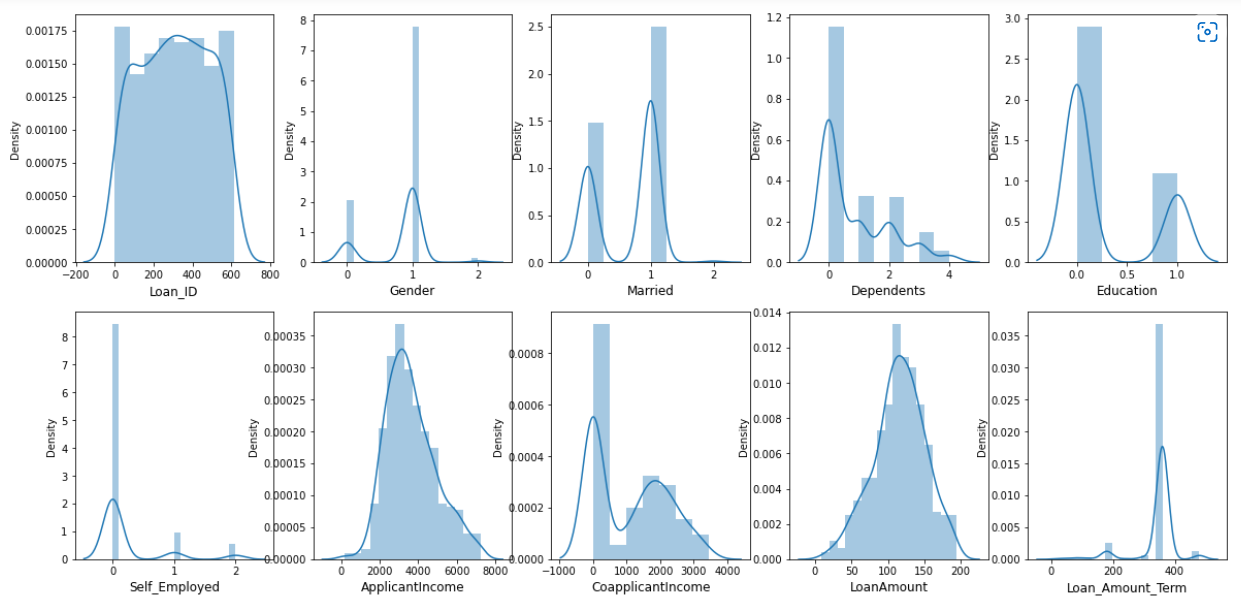


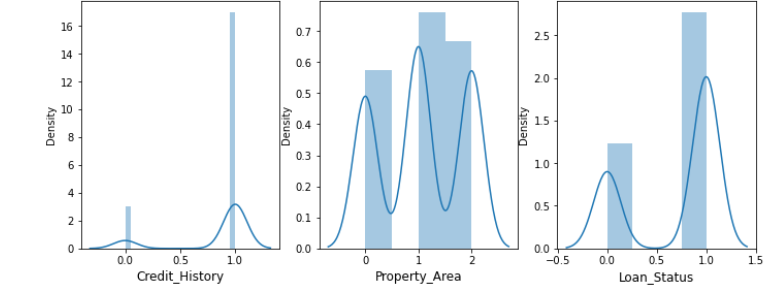


Like above I have removed for all the “Outliers” available.

After removing “Outliers”, below is the graph.







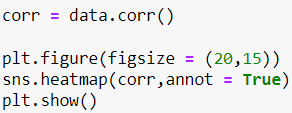
Now my dataset is equally distributed, and shape of data has been reduced to 440\*13.

**EDA Concluding Remark:**

1. I saw the object data type and it converted into numeric value.
2. I checked the null value present or not and it had so I filled these null values with the help of “mean” and “mode” method
3. I checked the total health of the dataset, and we found some deviation and high quantile range present in the dataset.
4. I checked the outliers present or not in the dataset with the help of “plotting” and it had so I removed these outliers with the help of “Quantile range” method.

**Pre-Processing Pipeline:**

Now I need to check the correlation between one feature to other features and plotted it with the help of heatmap. I do not want to consume any correlated features because correlated features consume extra space, and it will take more time, when I will build a model but there are not any correlated features available.



Chart, treemap chart

Description automatically generated

Now finally my dataset is totally cleaned.

Now I separated the features and label from the dataset.

A picture containing text

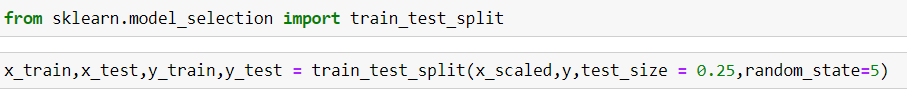
Description automatically generated

Next, I have scaled the features because I wanted to put my data in one scale and make it a standard value with the help of “standard scalar”

Graphical user interface, text, application, email

Description automatically generated

Now I divided my dataset into two parts one is train and second is test with the help of “train test split”. Generally, test size is 25 percent, but it’s totally depended on me how much we can contain the test size. In the training dataset I must train my dataset. When I will build the model, then remaining data I must test.

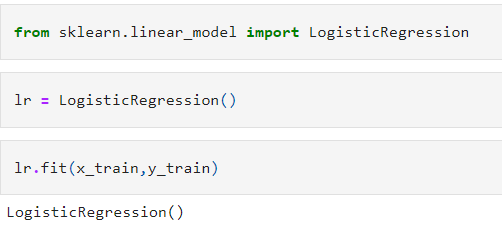


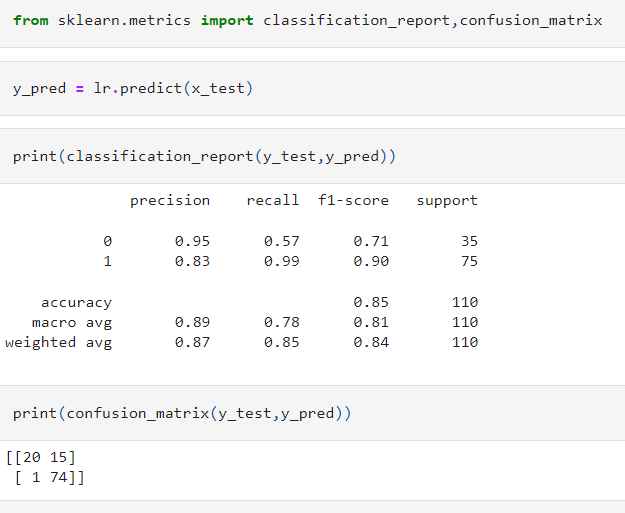
**Building Machine Learning Models:**

Now I am moving on building machine learning model.

**I have built four different machine learning mode.**

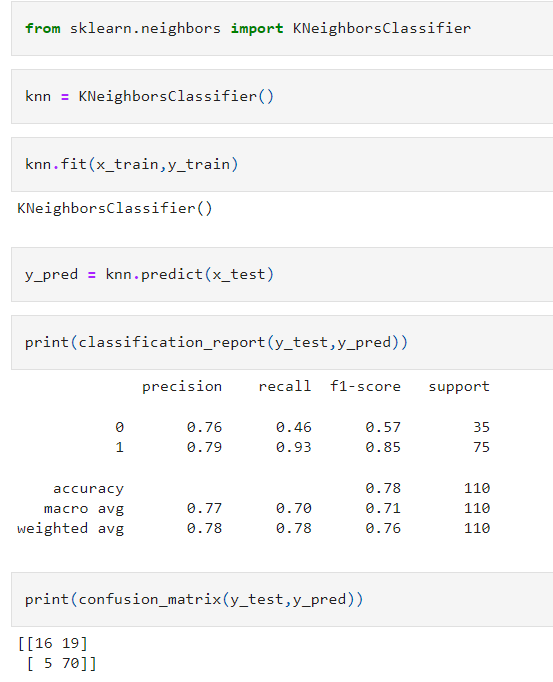
1. **Logistic regression model**





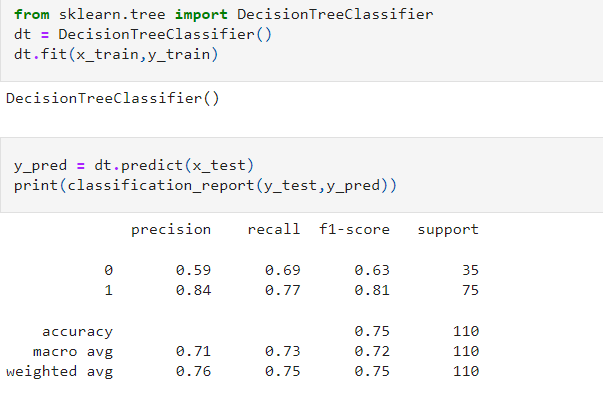
As we can see in above table, accuracy is 85%.

1. **Knn Classifier Model**



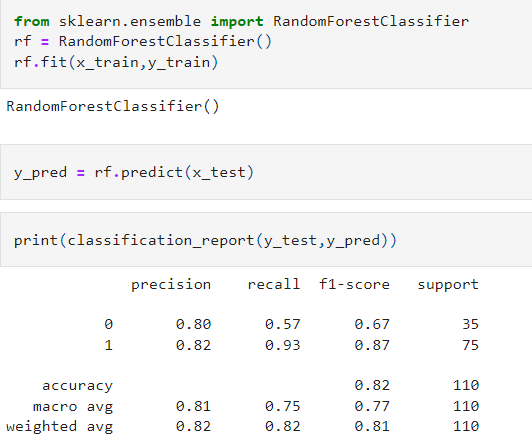
As we can see in above table, accuracy is 78%.

1. **Decision Tree Classifier Model**



As we can see in above table, accuracy is 75%.

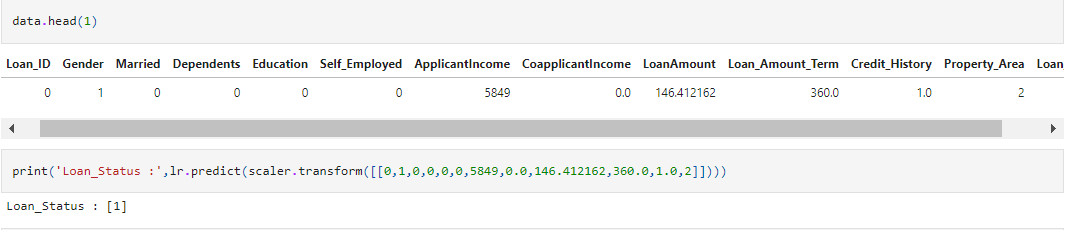
1. **Random Forest Classifier Model**



As we can see in above table, accuracy is 82%.

**Concluding Remarks**

In above all four model we can see, “ **Logistic regression model”** gives highest accuracy i.e 85% so we will be using this model for our project.



Next, we will tune our model with the help of “grid search cv”.

Graphical user interface, text, application, email

Description automatically generated

As we can after tuning also the accuracy is same.

Below is the confusion matrix for this model.

Graphical user interface, application, Word

Description automatically generated